Estimation and prediction of the HIV-AIDS-epidemic under conditions of HAART using mixtures of incubation time distributions

Heisterkamp, S. H.; de Vries, R.; Sprenger, H. G.; Hubben, G. A. A.; Postma, M. J.

Published in:
Statistics in Medicine

DOI:
10.1002/sim.2974

STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:781–794
Published online 27 June 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2974

Estimation and prediction of the HIV–AIDS-epidemic under conditions of HAART using mixtures of incubation time distributions

S. H. Heisterkamp[1, 2, *, †], R. de Vries[3], H. G. Sprenger[4], G. A. A. Hubben[3] and M. J. Postma[3]

[1]*Organon N.V., P.O. Box 20, Oss 5340 BH, The Netherlands*
[2]*Groningen Bioinformatics Centre, University of Groningen, The Netherlands*
[3]*Graduate Schools SHARE & GUIDE, Groningen Research Institute of Pharmacy, University of Groningen, Groningen, The Netherlands*
[4]*University Medical Centre Groningen (UMCG), Groningen, The Netherlands*

SUMMARY

The estimation of the HIV-AIDS epidemic by means of back-calculation (BC) has been difficult since the introduction of highly active anti-retroviral therapy (HAART) because the incubation time distributions needed for BC were poorly known. Moreover, it has been assumed that if the general public is aware that effective treatments are available then the majority of infected people would be known, and therefore a hidden epidemic was assumed not to exist. Nevertheless, it was suspected that not every infected person would come to the attention of health-care providers, and therefore estimates independent of the patients' registration were necessary. In this paper, the incubation time distributions for HIV treated with the HAART regimen are derived from a cohort study. By using estimates of the proportion treated according to the HAART regimen and the incubation time distributions estimated in the era before the implementation of HAART (pre-HAART), new marginal population incubation time distributions for each of the three risk groups (homosexuals, drug users and others) were constructed. The BC was performed using an empirical Bayesian approach based on the latter incubation time distribution. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: mixtures; back-calculation; empirical Bayes; incubation time distribution

1. INTRODUCTION

The use of back-calculation (BC) for estimating the number of unobserved HIV cases has long been an instrument of choice in the analysis of AIDS surveillance data. The crucial factor in the

*Correspondence to: S. H. Heisterkamp, Organon N.V., P.O. Box 20, Oss 5340 BH, The Netherlands.
†E-mail: siem.heisterkamp@organon.com

Received 17 April 2006
Accepted 8 May 2007

use of BC is an accurate description of the incubation time distribution either through a Markov model or through a survival function derived from cohort studies [1–3]. The last update of HIV incidence in Europe through BC occurred in 1996. Since the introduction of highly active anti-retroviral therapy (HAART) in 1997, the shape and form of the incubation time distribution have become uncertain, and prediction using BC only from AIDS-incidence data have been abandoned. Aalen *et al.* [4] proposed a Markov model in which both the reported HIV incidence (i.e. the treatment status of the patients) and AIDS incidence were taken into account, and therefore the HIV incidence can be modeled without the need to use the incubation time distribution.

It has been argued that since the publication of HAART, people at risk of HIV–AIDS might seek medical help at an earlier stage, such that virtually all HIV-infected people are registered. An effort was made to set up databases containing all known HIV-positive patients. In [5–7], an overview is given of the number of known HIV-diagnosed persons according to the year of diagnosis. However, this does not provide the incidence of HIV cases by year of onset or the number of unrecorded HIV infections.

Despite AIDS not being regarded as a threat in developed countries seven years after the introduction of HAART, there is still some concern. HAART does not provide complete protection, and because patients are required to adhere to a strict and complex drug regimen many become non-adherent. There are also some signs that the HIV prevention campaigns are not as effective as should be. Finally, the pharmaceutical industry simply wants to know how long and what quantity of medication will be required in future. For these reasons, our aim is to estimate the total number of HIV cases in the Netherlands by BC.

2. STUDY DESIGN

The cornerstone of the classic BC is the incubation time distribution, which reflects the incubation period from HIV to AIDS. Previously, in 1996, the incubation time distribution for untreated persons for the BC was derived from a Markov model, taking into account pre-AIDS death and going back and forwards in the defined stages of CD4 counts [1, 2]. CD4 counts were given as six categories, in decreasing numbers of CD4 cells per microliter. Group 1 being the group of patients with the highest count, $CD4 \geqslant 900$, group 2 with $900 > CD4 \geqslant 700$, group 3 $700 > CD4 \geqslant 500$, group 4 $500 > CD4 \geqslant 350$, group 5 $350 > CD4 \geqslant 200$ and group 6 $CD4 < 200$, respectively.

In the BC used for the European Union (EU) countries, the change of definition of AIDS and the dependency of the incubation period on the age of onset were taken into account [3]. For untreated persons, or persons not treated until the first HIV-positive test was performed, we still believe this to be the best incubation time distributions available for the EU countries. As these distributions were specifically adapted for each country using information on age distribution and date of change for the definition of AIDS, in this study we will use the specific distribution derived for the Netherlands. The incubation time distributions were all derived from disease progression data for stages defined by CD4 counts using a Markov model [1, 2]. In Figure 1, pathways for disease progression of an HIV-infected person are depicted. Note that HAART can occur only after 1996, thus making the incubation distribution time dependent on the calendar time of the first HIV test at which a decision for HAART was made.

For persons treated with anti-viral drugs based on their CD4 count, the estimation of the incubation time distribution becomes more complex. We divided the drug regimens into two different classes: HAART and pre-HAART. A regimen consisting of three or more different drugs

DOI: 10.1002/sim

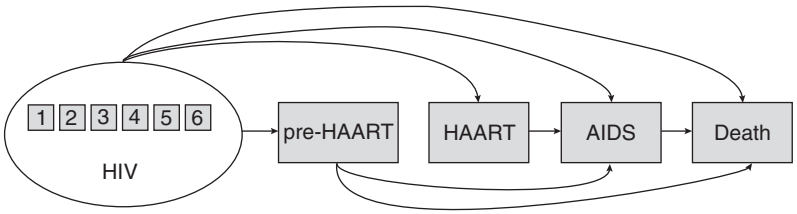

Figure 1. Markov model of the disease progression from HIV to death. The numbers 1–6 refer to CD4 stage at first HIV test (see text). Pathways between CD4 stages are described in [1, 2], in which pre-AIDS death is allowed, inclusion of age and change of AIDS definition in 1993 are described in [3].

Table I. Reported data adapted from [5–7].

Start	End	Homo	Drugs	Others	Total	At Risk (× 1000)
1978	1981	0	0	0	0	13 600
1981	1982	0	0	0	0	14 209
1982	1983	3	0	2	5	14 286
1983	1984	17	0	2	19	14 340
1984	1985	29	0	2	31	14 395
1985	1986	62	1	5	68	14 454
1986	1987	121	7	8	137	14 529
1987	1988	193	20	32	244	14 615
1988	1989	250	39	36	325	14 715
1989	1990	305	36	50	391	14 805
1990	1991	318	42	59	419	14 893
1991	1992	335	43	72	450	15 010
1992	1993	376	60	74	510	15 129
1993	1994	317	61	103	481	15 239
1994	1995	314	65	115	494	15 342
1995	1996	314	74	145	533	15 424
1996	1997	299	50	110	459	15 494
1997	1998	174	43	120	337	15 567
1998	1999	116	27	95	238	15 654
1999	2000	81	24	73	178	15 760
2000	2001	104	14	133	251	15 864
2001	2002	99	9	135	243	15 987
2002	2003	113	5	166	284	16 105
2003	2004	97	8	129	234	16 193
2004	2005					16 258

Note: At risk are the total population numbers of the Netherlands.

(e.g. two reverse transcriptase inhibitors and one protease inhibitor) was defined as HAART, while regimens consisting of two or less different drugs, as was usual before the implementation of HAART, were defined as pre-HAART. We used data from the HIV clinic at the University Medical Centre of Groningen (UMCG) to estimate the incubation time distribution while on treatment. The UMCG-HIV registration commenced in 1996 in anticipation that prospective and retrospective HIV research would be required to be performed in future. This observational research database comprises demographic data (e.g. age, gender and risk group), clinical data (e.g. HIV diagnosis, results of blood tests and disease classification) and resource utilization data (e.g. inpatient days

DOI: 10.1002/sim

and medication use). This registry is maintained by employees at the UMCG-HIV clinic. All persons in the registry with a first positive test were eligible for inclusion in this study. Patients were excluded if their medication history was not complete for the period they were in the registry. Patients were followed up till December 31, 2002, until they were removed from the registry or until death. A total of 552 patients were eligible for inclusion in the study. The following was recorded for each person: the CD4 count from the first positive test, the dates of starting and stopping medication, the date of being diagnosed with AIDS and the date of death. Dates were recorded in days and could be interval censored. Furthermore, we divided the patient population into three different risk groups: (i) 152 homosexual men; (ii) 37 drug users; and (iii) 202 others. The patients included in the risk group others mostly got infected by heterosexual transmission. AIDS incidence data for the Dutch situation were used as published and updated in [5–7] and are given in Table I.

3. INCUBATION TIME DISTRIBUTIONS

The incubation time distributions were estimated separately for the three risk groups and were derived through survival functions. The reason to fit incubation distributions for each risk group separately is that the course of the disease progression may depend on age, life style and other aspects of the health status, which may differ considerably between risk groups. For example, drug users have a greater risk for non-AIDS-related death, while the age distribution for the other risk groups may be different. The different stages and the possible pathways from HIV infected to AIDS or HIV infected to death, with the observed variations found in treatment regimes, are depicted in Figure 1. The data were analyzed using a parametric survival regression allowing for interval censoring (CensorReg from S-Plus, [8]). The best-fitted survival function appeared to be an extreme-value function (Gumbel). The stage defined by the CD4 count at the first positive HIV test was used as a categorical co-variate, while age and gender were ignored. For each risk group, the contribution of CD4 counts as well as the scale parameter were kept constant across the different pathways, while the intercept was allowed to vary. As we considered only the time from a certain treated (or untreated) stage to AIDS, which can be reached by several pathways through the different stages, we have to compute the convolution of the distributions associated with each of the nodes in a possible pathway. See [3] for details.

3.1. The marginal incubation time distribution from onset to AIDS

It is necessary to construct the marginal incubation time distribution for persons from onset to AIDS considering all possible pathways. For untreated persons and persons not treated until they are found to be sero-positive (with a recorded CD4 level), we used the incubation time distribution from the BC for the Dutch situation, which was published in 1996 [3].

In the classic BC model [9], the expected number of diagnosed AIDS cases y_i in some time interval $i = 1, \ldots, I$ is expressed by

$$E[y_i] = \sum_{j=1}^{J} P[x_i | z_j = 1] \cdot \beta_j N_j$$

where x_i is the event of being diagnosed with AIDS, $z_j \in \{0, 1\}$ the event of being infected in time interval $j = 1, \ldots, J$ and β_j the relative incidence of unobserved persons being infected in the

DOI: 10.1002/sim

later time interval and N_j the number at risk. We took for the number at risk the total population in the Netherlands in the subsequent years, simply as a means to have a common denominator. Usually one assumes for the distribution of y_i a Poisson distribution conditional on the known values of $P[x_i|z_j]$ and β_j. Also, a log-normal distribution of β_j is assumed with prior:

$$\log[\Delta_d(\beta|\ \beta_1 \ldots \beta_{j-d})] \propto \mathrm{N}(0, \lambda^{-1})$$

for some fixed d, where Δ_d is the difference operator of order d. Usually, it is sufficient to take first-order differences, i.e. $d=1$, which we call the neighbor prior. The aim of BC is to estimate β_j, i.e. the (partly) hidden epidemic curve. See [9] for details of the estimation problem. What can we say of $P[x_i|z_j]$ in the light of treatment effects? If no treatment would have been applied, we could safely use the discrete version of incubation time distribution used in earlier publications. However, since at some time $j=t_0$ pre-HAART and HAART treatments have been used, this must be taken into account. We may write in general the convolution:

$$P[x_i|z_j=1]=\sum_{k=1}^{6}\sum_{t=j}^{i} P[\mathrm{CD4}_{t-i}=k|z_j=1]P[x_i|z_j=1\wedge \mathrm{CD4}_{t-i}=k]$$

where $\mathrm{CD4}_{t-i}=k$ is the event of an untreated person being in the kth class of the CD4 classes ($k=1,\ldots,6$) at time period $t-i$. The moment a person gets treated we have to change the probability of getting AIDS in the last term of the summation.

Now, from $j=t_0$ on, there is a chance that an infected person will get on treatment with pre-HAART or HAART. Assume that there exist non-zero probabilities $\pi_{skl} s=t_0,\ldots,J$ that a person starts treatment l when arriving at a certain class k of CD4 counts at time s. We may then derive the following equality:

$$P[x_i|z_j=1\wedge \mathrm{CD4}_{t-i}=k]=\sum_{l}\pi_{t-ikl}P[x_i|z_j=1\wedge \mathrm{CD4}_{t-i}=k\wedge I_{\text{treatment}=l}]$$

The treatment indicator is in the case of three treatments, untreated (0), pre-HAART (1) and HAART (2), while in the case of two treatments we drop the category pre-HAART.

However, as we already have the 'old' incubation time distribution for untreated persons, this equation is not used for untreated persons. Thus, given the probabilities $\pi_{skl}, s=t_0,\ldots,J$, we have to derive the probability of getting AIDS in time period i when infected at a certain time j using the 'old' incubation time distributions, together with the prevalence of being in class k of the CD4 counts derived from these, and the new extreme value distributions for each pathway

$$\begin{aligned}
&P[x_i|z_j=1]\\
&=\sum_{k}\sum_{t=j}^{\min(t_0-1,i)} P[\mathrm{CD4}_{t-i}=k|z_j=1]P[x_i|z_j=1\wedge \mathrm{CD4}_{t-i}=k\wedge I_{\text{treatment}=0}]\\
&\quad+\sum_{k}\sum_{t=\min(t_0,i)}^{i} P[\mathrm{CD4}_{t-i}=k|z_j=1]\sum_{l=0}^{2}\pi_{t-ikl}P[x_i|z_j=1\wedge \mathrm{CD4}_{t-i}=k\wedge I_{\text{treatment}=l}]
\end{aligned}$$

DOI: 10.1002/sim

We write the above equation in a form that enables us to discriminate between the untreated and treated persons

$$
\begin{aligned}
&P[x_i|z_j=1]\\
&=\sum_k \sum_{t=j}^{\min(t_0-1,i)} P[\mathrm{CD4}_{t-i}=k|z_j=1]P[x_i|z_j=1\wedge \mathrm{CD4}_{t-i}=k\wedge I_{\text{treatment}=0}]\\
&+\sum_k \sum_{t=\min(t_0,i)}^{i} \pi_{t-ik0}P[\mathrm{CD4}_{t-i}=k|z_j=1]P[x_i|z_j=1\wedge \mathrm{CD4}_{t-i}=k\wedge I_{\text{treatment}=0}]\\
&+\sum_k \sum_{t=\min(t_0,i)}^{i} P[\mathrm{CD4}_{t-i}=k|z_j=1]\sum_{l=1}^{2}\pi_{t-ikl}P[x_i|z_j=1\wedge \mathrm{CD4}_{t-i}=k\wedge I_{\text{treatment}=l}]
\end{aligned}
$$

It is clear that only for $i<t_0$ the incubation incidence is simply the untreated one. For $i\leqslant t_0$ it is not possible to use the untreated incubation distribution directly as for the untreated persons we have to down-weight the tail of the distribution by the probability of going onto treatment.

As the incubation time distribution for the untreated persons is not available in the fine details needed, we used the following approximation to the untreated part of the incubation time distribution:

$$
P[x_i|z_j=1\wedge I_{\text{treatment}=0}]=w_{i,j}P_0[x_i|z_j=1]w_{i,j}\begin{cases}1, & i<t_0\wedge j<t_0\\ \bar{\pi}_0, & i\geqslant t_0\wedge j\geqslant t_0\\ \bar{\pi}_0 e^{\kappa(j-i)}, & i\geqslant t_0\wedge j<t_0\end{cases}
$$

where $P_0[\cdot]$ is the original distribution function, $\bar{\pi}_0$ the average proportion not on treatment and κ some tuning constant. Thus, we down-weight the original distribution heavier for those who are infected before the treatment regimes started. Constraints are applied on the weights such that the weights and the mean probabilities of getting treatment sum to 1.

It may be tempting to try to estimate the probabilities π_{tkl} simultaneously with β_j as we have a finite mixture of distributions. However, when fitted with a fixed arbitrary π_{tkl}, it is clear that a fit with $\bar{\beta}_j=\pi_{tkl}\beta_j$ will fit equally well and thus cancel out the other treated components, or *vice versa*. Unless there is explicit information on the AIDS-diagnosed patients with respect to their treatment status, a simultaneous fit is impossible. Thus, we estimated the proportions on treatment from the Groningen cohort by a simple generalized linear model for each risk group. From this we concluded that—after 1999—the proportion on treatment remained constant and apart from the risk-group 'others', were independent of CD4 class. We chose to ignore the latter dependance as the numbers on which these estimates were based were small. The estimated treatment proportions increased from 0.11 in 1996 to 0.45 in 1999 and we used the indices of the treatment proportion depending linearly on time from 1996 onwards until 2000. In Figure 2, the marginal distributions are given on the basis of the selected infection cohorts for homosexual men.

DOI: 10.1002/sim

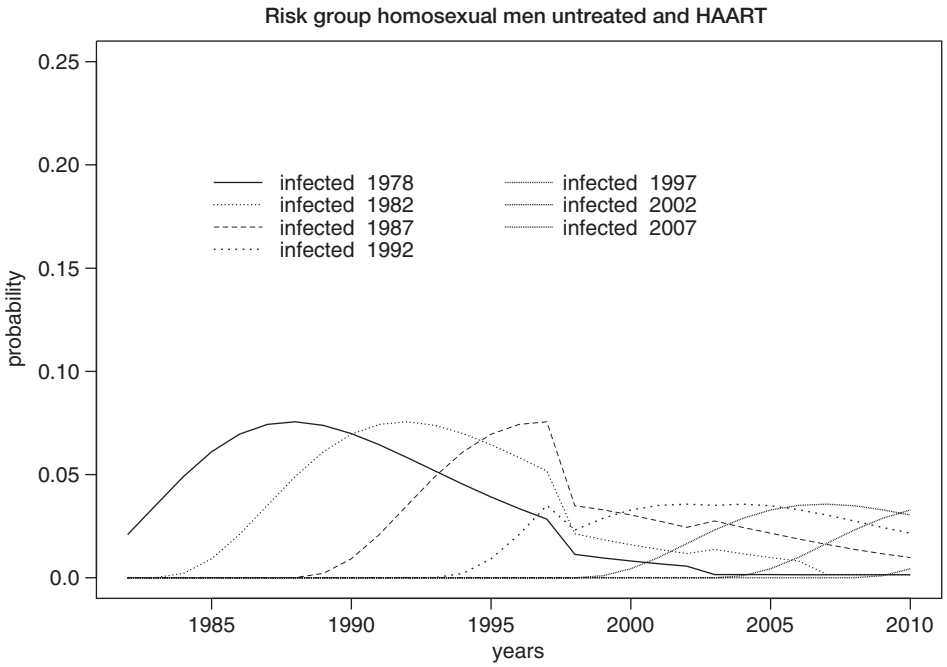

Figure 2. Incubation time distributions for selected infection cohorts of homosexual men.

4. FITTED BACK-CALCULATION MODELS

For each separate risk group, a BC model was fitted to estimate the relative incidence of HIV, using the number of inhabitants as the number of people at risk, i.e. a log-link with the log of the number at risk as offset. We tried marginal incubation time distributions based on two- and three-treatment classes, and subsequently used the Akaike Information Criterion (AIC) as the goodness-of-fit measure. For all of the risk groups—except drug users—the two-treatment marginal incubation time distribution gave by far the smallest AIC. Only the risk group of drug users fitted much better with three-treatment classes. Figures 2–8 depict for each risk group (i) the fitted AIDS incidence in numbers and (ii) the relative HIV incidence (per 1000), all with the accompanying predictions from 2005 to 2010. The figures were created using incubation time distribution based on two-treatment classes for all risk groups. The 95 per cent prediction limits are based on the use of the negative binomial distribution, details of which can be found in [9]. Note that the predictions for the relative HIV incidence after 2003 are simply the last carry forward of the estimate in 2003 as a result of the neighbor prior in the Bayesian prediction model.

4.1. Prevalence and cumulative incidence of HIV in the Netherlands

We estimated for each year (i) the AIDS incidence with their prediction intervals, (ii) the HIV incidence and prevalence with their prediction intervals, (iii) death incidence and (iv) the cumulative numbers of HIV cases and deceased for each risk group. For reasons of brevity, we include in Table II only HIV prevalence and cumulative incidence for the three risk groups. In particular,

DOI: 10.1002/sim

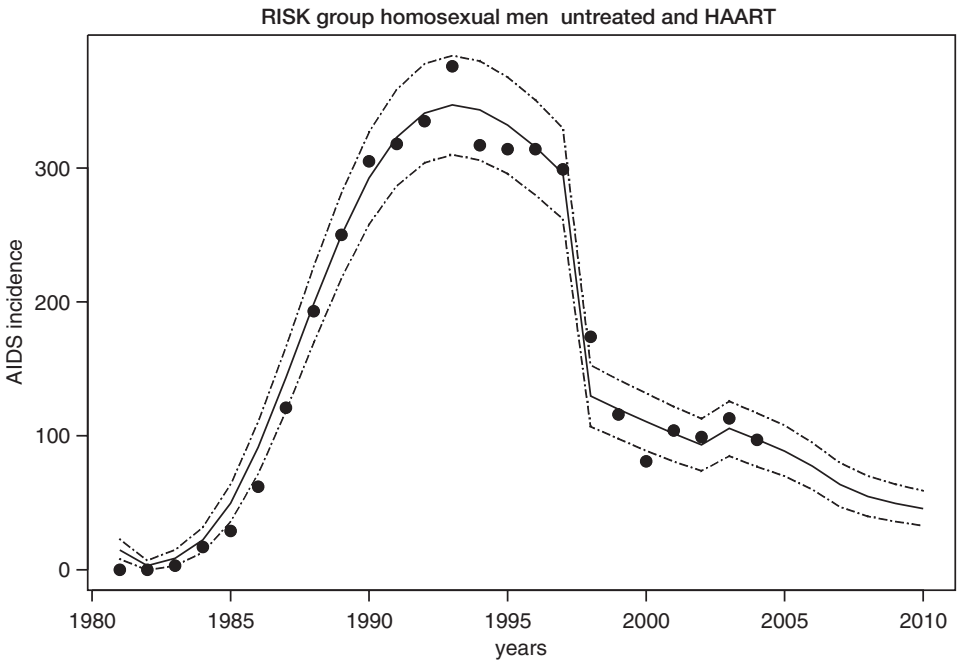

Figure 3. AIDS incidence for homosexual men.

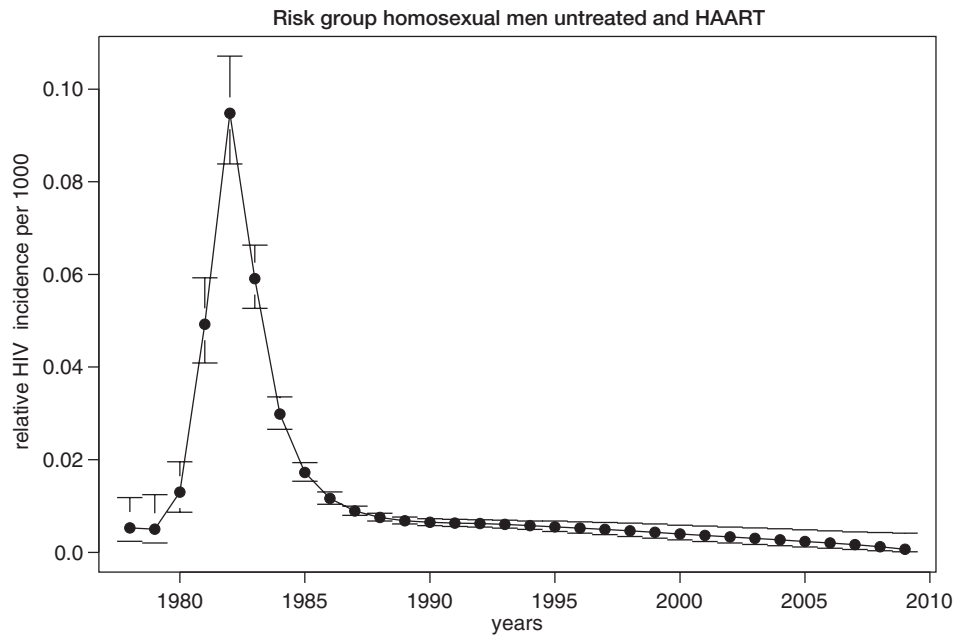

Figure 4. Relative HIV incidence per 1000 for homosexual men.

DOI: 10.1002/sim

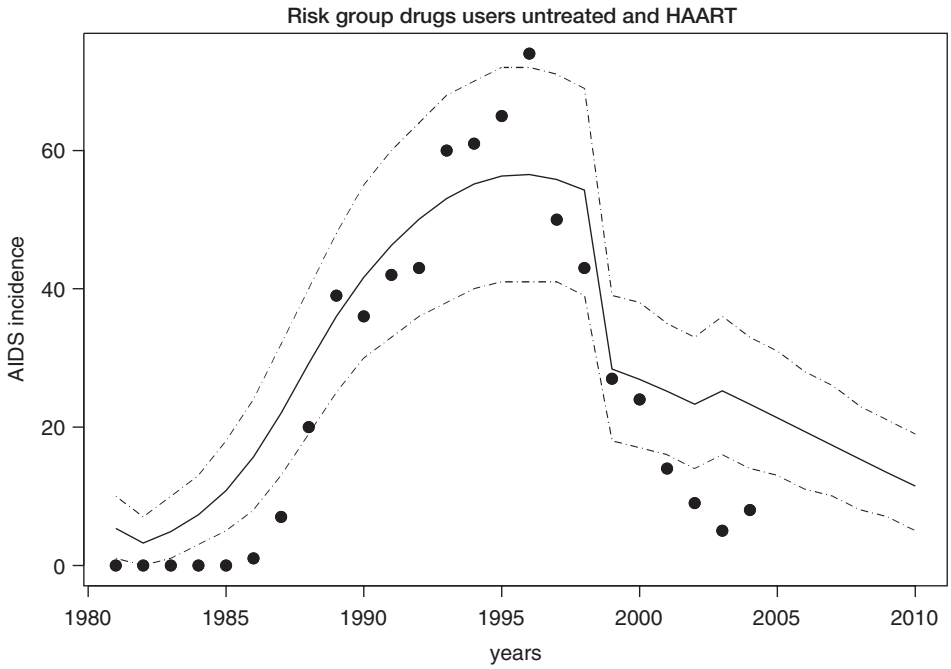

Figure 5. AIDS incidence for drugs users.

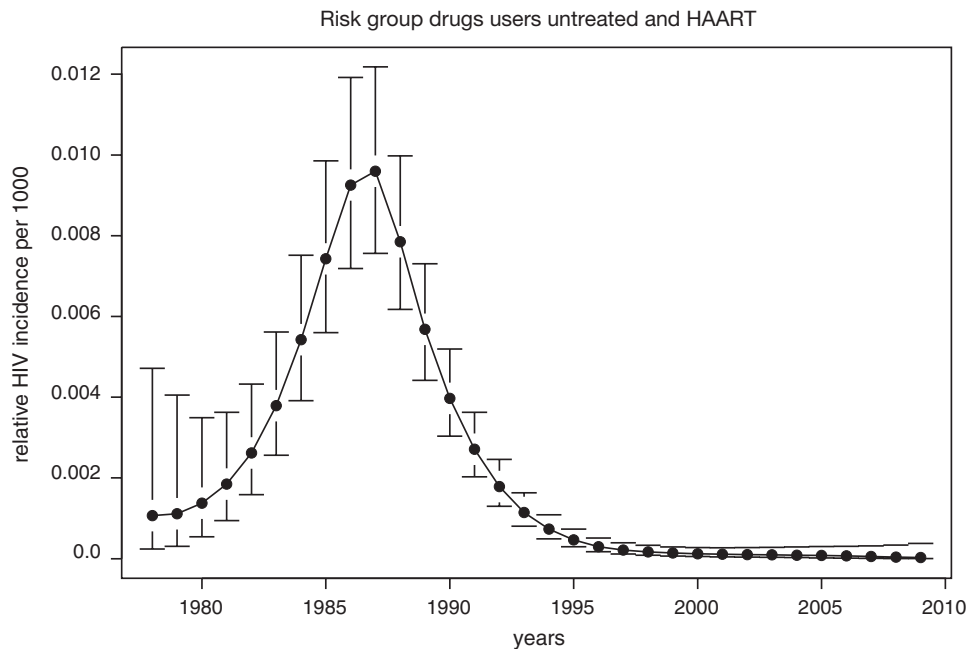

Figure 6. Relative HIV incidence per 1000 for drugs users.

DOI: 10.1002/sim

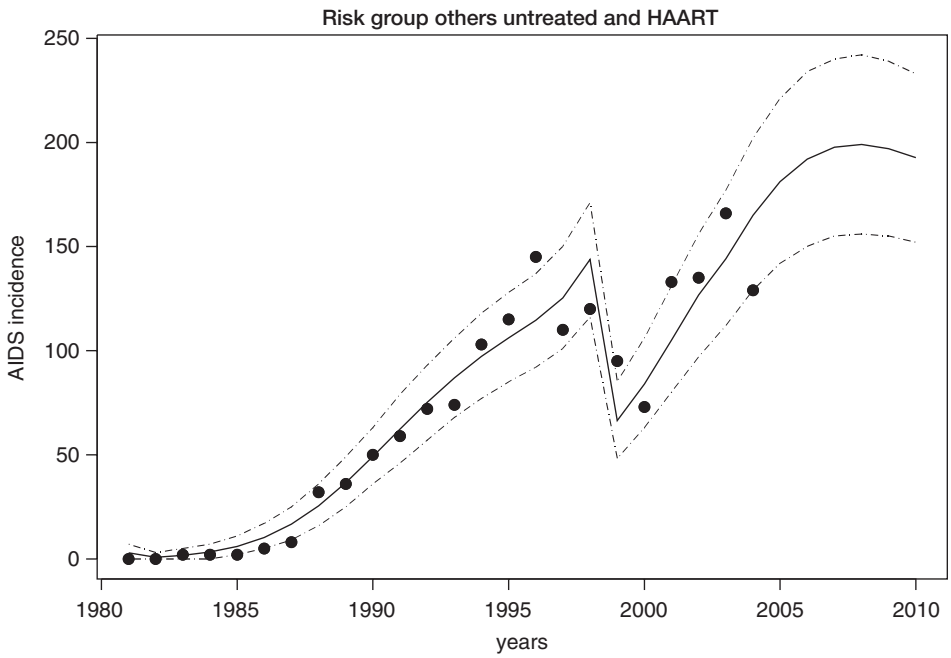

Figure 7. AIDS incidence for risk group 'others'.

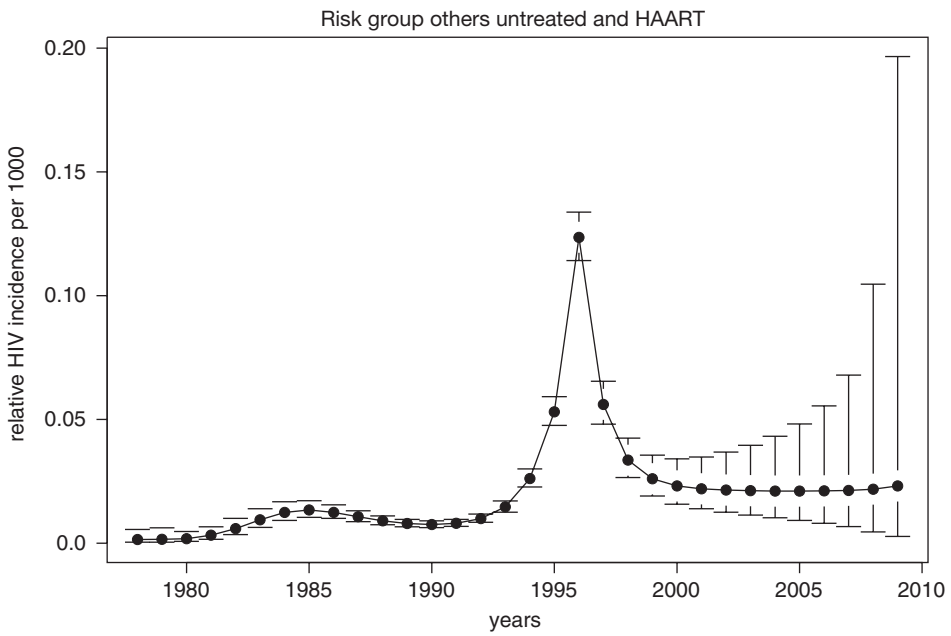

Figure 8. Relative HIV incidence per 1000 and prediction for risk group 'others'.

DOI: 10.1002/sim

Table II. Estimated HIV-cumulative incidence and HIV incidence in absolute numbers for homosexual men, drugs users and risk group others.

	Homosexual men		Drugs users		Others	
Year	HIV cumulative	Prevalence	HIV cumulative	Prevalence	HIV cumulative	Prevalence
1981	216	197	1	1	59	55
1982	286	262	1	1	81	75
1983	472	436	3	2	106	98
1984	1178	1115	7	5	15	139
1985	2542	2427	19	13	236	216
1986	3396	3191	45	32	372	340
1987	3830	3487	97	72	553	502
1988	4082	3553	197	147	748	672
1989	4253	3492	386	303	931	820
1990	4385	3355	683	563	1089	931
1991	4498	3172	806	644	1223	1007
1992	4600	2964	868	661	1342	1056
1993	4698	2746	917	660	1456	1091
1994	4794	2530	962	653	1578	1124
1995	4889	2324	995	631	1730	1178
1996	4982	2131	1014	595	1956	1298
1997	5072	1952	1024	550	2360	1588
1998	5158	1920	1031	503	3186	2280
1999	5239	1892	1035	495	5121	4153
2000	5317	1870	1038	486	6004	4957
2001	5391	1851	1041	479	6536	5389
2002	5460	1835	1045	472	6952	5685
2003	5524	1803	1048	464	7324	5920
2004	5583	1774	1052	458	7680	6118
2005	5637	1747	1057	452	8029	6293
2006	5685	1724	1062	448	8373	6453
2007	5729	1708	1069	445	8715	6604
2008	5767	1695	1076	444	9057	6753
2009	5800	1681	1085	445	9399	6905
2010	5827	1665	1095	448	9747	7064

the cumulative number of HIV cases is interesting. At the beginning of the epidemic in 1987, the cumulative cases of HIV-infected persons was estimated to be 12 000. This estimate was lowered to 9 000 in 1995 [3]. The cumulative number of recorded HIV persons as of June 2005 is, according to [7], 5556 for homosexual men, 563 for drug users and 4500 for the risk group others. From Table II, our estimates for the cumulative incidence by the end of 2004 are 5583 for the risk group of homosexual men, 1052 for drug users and 7680 for the risk group others. Allowing for a linear interpolation between the estimates for the end of 2004 and 2005, our estimates until June 2005 are: 5605, 1053 and 7825, respectively. Thus, there is a very good agreement in our estimates for the risk group homosexual men and that of the UMCG database. This implies that homosexual men are more aware of their risk for HIV/AIDS. The discrepancy for the drug users is most likely due to our modeling technique. Specifically, our pre-HAART model assumed that there was an extra 15 per cent of deaths before AIDS (and probably HIV) was diagnosed [3]. In addition, pre-HAART drug users accounted for the majority of the pre-HAART incidence in the total population. If our

DOI: 10.1002/sim

incubation time distribution is correct, there are approximately 3300 HIV-infected persons missing from the risk group others, which is likely due to mostly heterosexual transmission. It is clear that for the risk groups homosexual men and intravenous drug users the epidemic is over. The risk group others (which comprises mostly of heterosexuals) had a maximum incidence in 1995, which coincides with the EU report in 1996. After the peak in 1995, the incidence has decreased, but is still at a higher level than prior to 1993 and at a much higher level than in the homosexual men and drug user groups. It is of concern to the public health authorities in the Netherlands that 43 per cent of the risk group others is not registered in the database and therefore they are probably unaware of their HIV status. We estimate that by the end of 2004 the prevalence among the risk group others is 6118 (95 per cent prediction interval: 5475–6835). While the prevalence among the risk group of homosexual men is 1774 (1668–1886), intravenous drug users account for 485 (288–729) persons. Thus, we estimate that from the total HIV-infected population, 73 per cent belongs to the risk group others, in contrast to the reported estimate of 43 per cent in [7].

5. DISCUSSION AND CONCLUSION

We have derived a method of estimating the total incidence of HIV from publicly available AIDS-incidence data, without the need for detailed information about each individual's treatment history. The incubation time distributions, which are needed for classic BC, were derived by means of both the known distributions for untreated persons used in [3] (available from the first author for other EU countries as S-Plus files), and fitted distributions from a local cohort, together with information on the proportion of persons going on different treatment regimes. Although the possibility exists that the incubation time distributions derived for treated persons have a local component because treatments are adapted to the local population locally, these incubation time distributions could be used by other countries in the absence of other information. Thus, the only truly local parameters are the proportions of persons commencing treatment as defined above, which can be simply estimated from hospital-based data. We therefore believe that the derived methodology can be applied to other EU countries as well. From the first author's incubation time distributions for the EU countries as used in [3], the parameters used in the extreme value distributions as well as S-Plus software for BC are available. The results of our study suggest that for the classic risk groups (homosexual men and intravenous drug users) there is evidence that the recorded HIV infections are well covered. This is probably due to the thorough follow-up of patients by the health-care professionals, awareness created by the health authorities and the Dutch government's policy regarding the use of hard drugs. However, it is apparent that heterosexuals who form the majority of the others risk group are receiving little attention. In fact, we estimate that 43 per cent of the others group is not registered and, therefore, they are probably unaware of their infection status. Moreover, we estimate that the others account for a much higher (73 per cent) percentage of the total infected population than the recently reported 43 per cent in [7].

5.1. Convolution of extreme value distributions

In this section, the distribution of the sum of independently distributed extreme value distributions with common-scale parameter is derived. Let $X_i,\ i = 1, \ldots, k$, be independent stochastic variables with the extreme value distribution with location parameters $\alpha,\ i = 1, \ldots, k$, and the common-scale

DOI: 10.1002/sim

parameter β:

$$F_{X_i}(x, \alpha_i, \beta) = e^{-e^{-(x-\alpha_i)}/\beta} \quad -\infty<\alpha_i<\infty, \quad \beta>0, \quad i=1, \ldots, k \tag{1}$$

The mean and variance of this distribution are $\alpha_i + 0.577216\,\beta$ and $\pi^2\beta^2/6$, respectively, see [10]. We seek the cumulative distribution of $Z = \sum_i X_i$. The moment-generating function of (1) is given by

$$M_{X_i}(t) = e^{\alpha_i t}\Gamma(1-\beta t) \tag{2}$$

From standard theory, it follows that the moment-generating function of $\sum_i X_i$ equals:

$$\begin{aligned} M_Z(t) &= e^{\sum_i \alpha_i t}(\Gamma(1-\beta t))^k \quad \text{with } Z = \sum_i X_i \\ &= (e^{t\bar{\alpha}}\Gamma(1-\beta t))^k \quad \text{with } \bar{\alpha} = \sum_i \alpha_i \end{aligned} \tag{3}$$

From this, it follows that the distribution of Z is given by the convolution of k independent identically distributed Gumbel distributions given by (1) with parameters $\bar{\alpha}$ and β. It is in principle possible to find the density function by a complex integration from (3), but the form looks intractable for convolutions $k>2$. Numerical comparison reveals that for k not too large the form of the distribution will remain Gumbel. Thus, by letting the first two moments of Z, assuming a Gumbel distribution, coincide with the first two moments of the sum of identically distributed Gumbel distributions with parameters $\bar{\alpha}$ and β, we find that the parameters of the corresponding Gumbel distribution of Z are approximately:

$$\begin{aligned} \tilde{\alpha} &= \sum_i \alpha_i + 0.577216 \cdot \beta \cdot k(1 - 1/\sqrt{k}) \\ \tilde{\beta} &= \beta\sqrt{k} \end{aligned}$$

However, for $k=2$, we may derive directly from the convolution of two i.i.d. Gumbel distributions the corresponding distribution. Thus, let $Z = X + Y$, with Z and Y independently distributed. Then, the c.d.f. of Z is given by the convolution:

$$F_Z(z) = \int_{-\infty}^{\infty} F_Y(z-x) f_X(x)\,\mathrm{d}x$$

where F_Y and f_x denote the c.d.f. of Y and the p.d.f. of Y, respectively. The p.d.f. of the Gumbel distribution is

$$f_X(x, \alpha, \beta) = \frac{1}{\beta} e^{-e^{-(x-\alpha)/\beta}} e^{-(x-\alpha)/\beta}$$

From this, it follows that the convolution of two i.i.d. Gumbel distributions with parameters α and β is

$$F_{X+X}(z) = \frac{1}{\beta}\int_{-\infty}^{\infty} e^{-e^{-(z-x-\alpha)/\beta}} e^{-e^{-(x-\alpha)/\beta}} e^{-(x-\alpha)\beta}\,\mathrm{d}x$$

DOI: 10.1002/sim

$$= \frac{1}{\beta} \int_{-\infty}^{\infty} \mathrm{e}^{-\mathrm{e}^{-(x-\alpha)/\beta(1+\mathrm{e}^{-(z-2\alpha)/\beta})}} \mathrm{e}^{-(x-\alpha)/\beta} \, \mathrm{d}x$$

$$= (1 + \mathrm{e}^{-(z-2\alpha)/\beta})^{-1}$$

By a change of variables, i.e. $w = \mathrm{e}^{-(z-2\alpha)}/\beta$, and thus translating the time z into $z = 2\alpha - \beta \log(w)$, we see that w has exactly the standard logistic distribution with mean 0 and variance $\pi^2/3$.

ACKNOWLEDGEMENTS

This study was conducted at the time when the first author was employed by the National Institute for Public Health and the Environment, Bilthoven, The Netherlands.

REFERENCES

1. Hendriks JCM, Satten GA, van Ameijden EJ, van Druten JAM, Coutinho RA, van Griensven GJP. The incubation period of AIDS in injecting drug users estimated from prevalent cohort data, accounting for death prior to an AIDS diagnosis. *AIDS* 1998; **12**:1537–1544.
2. Hendriks JCM, Satten GA, Longini IM *et al*. Use of immunological markers and continuous-time Markov models to estimate progression of HIV infection in homosexual men. *AIDS* 1996; **10**:649–656.
3. Downs AM, Heisterkamp SH, Rava L, Houweling H, Jager CJ, Hamers F. Back-calculation by birth cohort, incorporating age-specific disease progression, pre-AIDS mortality and change in European AIDS case definition. *AIDS* 2000; **14**:2179–2189.
4. Aalen OO, Farewell VT, De Angelis D, Day NE, Gill ON. A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: application to AIDS prediction in England and Wales. *Statistics in Medicine* 1997; **16**:2191–2210.
5. Op de Coul ELM, van Valkengoed IGM, van Sighem AI, de Wolf F, van de Laar MJW. *RIVM Report 441100018/2003*, Bilthoven, 2003.
6. Op de Coul ELM, van Valkengoed IGM, van Sighem AI, de Wolf F, van de Laar MJW. *RIVM Report 441100018/2004*, Bilthoven, 2004.
7. van Laar MJW, de Boer IM, Koedijk FDH, Op de Coul ELM. *RIVM Report 441100022/2005*, Bilthoven, 2005.
8. Data Analysis Division. *S-Plus 6.0 Guide to Statistics*, vol. 2. MathSoft, Seattle, WA, 2000.
9. Heisterkamp SH, van Houwelingen JC, Downs AM. Empirical Bayesian estimators for a poisson process propagated in time. *Biometrical Journal* 1999; **41**(4):385–400.
10. Mood AM, Graybill FA, Boes DC. *Introduction to the Theory of Statistics*. McGraw-Hill: New York, 1974.

DOI: 10.1002/sim